

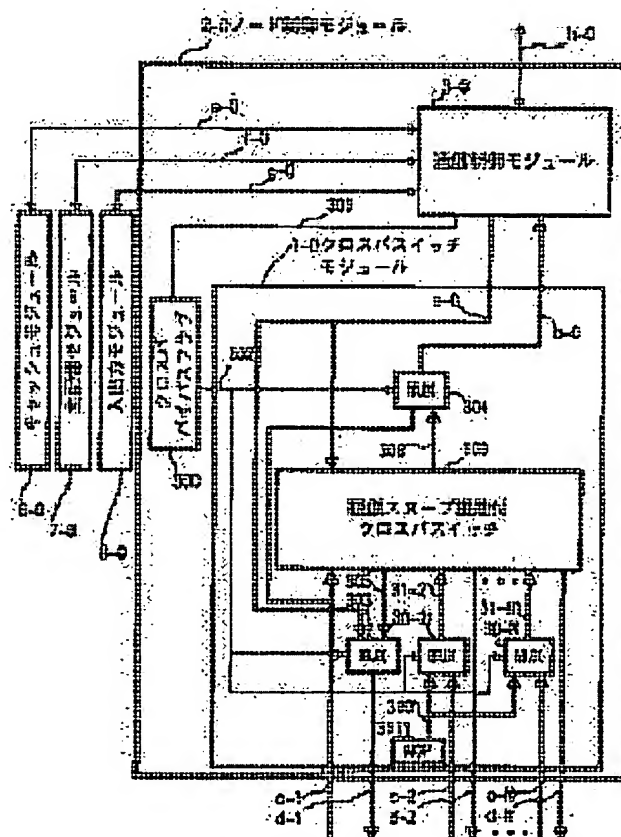


MAIN STORAGE SHARING TYPE MULTIPROCESSOR SYSTEM**Publication number:** JP2000348000 (A)**Publication date:** 2000-12-15**Inventor(s):** TANAKA TAKESHI; AKASHI HIDEYA; TSUSHIMA YUJI;
UEHARA KEITARO; HAMANAKA NAOKI; SHONAI TORU**Applicant(s):** HITACHI LTD**Classification:****- international:** G06F15/173; G06F12/00; G06F12/08; G06F15/16;
G06F15/177; G06F15/16; G06F12/00; G06F12/08; (IPC1-
7): G06F15/16; G06F15/173; G06F15/177**- European:** G06F12/08B4N**Application number:** JP19990156560 19990603**Priority number(s):** JP19990156560 19990603**Also published as:** JP3721283 (B2) US6789173 (B1)**Abstract of JP 2000348000 (A)**

PROBLEM TO BE SOLVED: To reduce the cost of development for a main storage sharing type multiprocessor system by using a common node between a small scale system of a small number of nodes and a large scale system of a large number of nodes. **SOLUTION:** Each of nodes constructing a multiprocessor has a constitution shown in a diagram and also has plural processors having caches which are connected to a bus (h). A crossbar switch module 1-0 of a node control module 2-0 has a crossbar switch 100 having a pseudo snoop function having a cache coherency control function. A crossbar bypass flag 300 controls an MUX and transfers the information to other nodes other than its own one with no intervention of the switch 100 when the flag is set at 1 to invalidate the switch 100.; Then the information is transferred to other nodes via the switch 100 when the flag is set at 0 to validate the switch 100. An outside crossbar switch is placed outside the nodes and all node flags are set at 1 when the nodes are connected to the outside crossbar switch.



Data supplied from the esp@cenet database — Worldwide

Family list2 application(s) for: **JP2000348000 (A)****1 MAIN STORAGE SHARING TYPE MULTIPROCESSOR SYSTEM****Inventor:** TANAKA TAKESHI ; AKASHI HIDEYA **Applicant:** HITACHI LTD
(+4)**EC:** G06F12/08B4N**IPC:** G06F15/173; G06F12/00; G06F12/08; (+8)**Publication info:** JP2000348000 (A) — 2000-12-15

JP3721283 (B2) — 2005-11-30

2 Node controller for performing cache coherence control and memory-shared multiprocessor system**Inventor:** TANAKA TSUYOSHI [JP] ; AKASHI HIDEYA [JP] (+4) **Applicant:** HITACHI LTD [JP]**EC:** G06F12/08B4N**IPC:** G06F15/173; G06F12/00; G06F12/08; (+6)**Publication info:** US6789173 (B1) — 2004-09-07

Data supplied from the esp@cenet database — Worldwide

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2000-348000

(P2000-348000A)

(43)公開日 平成12年12月15日(2000.12.15)

(51)Int.Cl. ⁷	識別記号	F I	テ-マコ-ト*(参考)
G 0 6 F 15/16	6 4 5	G 0 6 F 15/16	6 4 5 5 B 0 4 5
15/173		15/173	C
15/177	6 8 2	15/177	6 8 2 J

審査請求 未請求 請求項の数 5 O L (全 12 頁)

(21)出願番号 特願平11-156560

(22)出願日 平成11年6月3日(1999.6.3)

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 田中 剛

東京都国分寺市東荏ケ窪一丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 明石 英也

東京都国分寺市東荏ケ窪一丁目280番地

株式会社日立製作所中央研究所内

(74)代理人 100099298

弁理士 伊藤 修 (外1名)

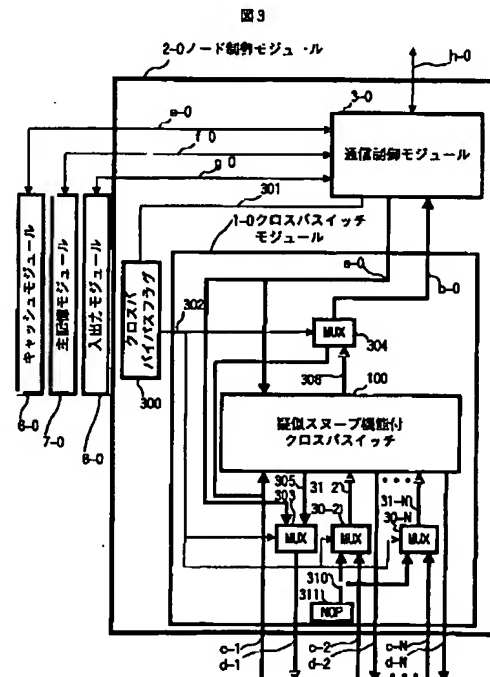
最終頁に続く

(54)【発明の名称】 主記憶共有型マルチプロセッサシステム

(57)【要約】

【課題】 ノード数の少ない小規模システムとノード数の多い大規模システムで共通のノードを使うことにより開発コストを削減することにある。

【解決手段】 マルチプロセッサを構成する各ノードは図に示す構成にさらにバスにキャッシュを有するプロセッサを複数台接続した構成を有する。ノード制御モジュール2-0のクロスバスイッチモジュール1-0はキャッシュコヒーレンシ制御機能を持つ疑似スヌープ機能付クロスバスイッチ100を有し、クロスババイパスフラグ300はMUXを制御し、フラグが1のとき他のノード間との情報転送は該スイッチ100を介さず行われ、該スイッチ100は無効になり、フラグが0のとき他のノード間との情報転送は該スイッチ100を介して行われ、該スイッチ100は有効となる。ノード外部に外部クロスバスイッチを設け、各ノードを該スイッチに接続する場合には、全ノードのフラグを全て1にする。



【特許請求の範囲】

【請求項1】 複数のノードを有する主記憶共有型マルチプロセッサシステムにおいて、

該各ノードは、キャッシュメモリを有する1個以上のCPUモジュールと、1個以上の主記憶モジュールと、これらのモジュールと他ノードの間の通信制御を行うノード制御モジュールを備え、

該ノード制御モジュールは、ノード間通信のインタフェースを制御する通信制御モジュールと、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンス制御を行うクロスバスイッチを有するクロスバスイッチモジュールを有し、

該クロスバスイッチモジュールは、モードレジスタを有し、該モードレジスタにセットされた値に応じて該クロスバスイッチを有効、あるいは無効にすることを特徴とする主記憶共有型マルチプロセッサシステム。

【請求項2】 請求項1記載の主記憶共有型マルチプロセッサシステムにおいて、

前記クロスバスイッチモジュールは、前記モードレジスタに代えてモード信号ピンを有し、該モード信号ピンの信号値に応じて該クロスバスイッチを有効、あるいは無効にすることを特徴とする主記憶共有型マルチプロセッサシステム。

【請求項3】 請求項1または請求項2記載の主記憶共有型マルチプロセッサシステムにおいて、

前記クロスバスイッチモジュールは、全ノードから前記クロスバスイッチに転送されるメモリアクセス要求がどのノード内の主記憶モジュールに割り付けられているアドレス空間に対する要求であるかを判別する手段を有することを特徴とする主記憶共有型マルチプロセッサシステム。

【請求項4】 請求項1乃至請求項3のいずれかの請求項記載の主記憶共有型マルチプロセッサシステムにおいて、

前記ノードの外部に、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンス制御を行うクロスバスイッチを有する外部クロスバスイッチモジュールを設け、

前記各ノードを該外部クロスバスイッチモジュールに接続し、

前記各ノード内のクロスバスイッチを無効にすることを特徴とする主記憶共有型マルチプロセッサシステム。

【請求項5】 請求項1乃至請求項3のいずれかの請求項記載の主記憶共有型マルチプロセッサシステムにおいて、

前記ノード間を直接接続し、少なくとも1つのノード内のクロスバスイッチを有効にすることを特徴とする主記憶共有型マルチプロセッサシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、一般にプロセッサが主記憶を共有し、キャッシュメモリのコヒーレンシ制御のために、要求キャッシュラインのアドレスを全プロセッサに配布するようなスヌープ方式を使用しているマルチプロセッサシステムに係り、特に、CPUと主記憶、及びキャッシュメモリユニットが搭載され、小規模システムと大規模システムで共通の部品として使用できる構成のノードを有するマルチプロセッサシステムに関する。

【0002】

【従来の技術】従来技術として、CPUモジュールと主記憶モジュールを同一のボードに搭載したノードを、小規模マルチプロセッサシステムと大規模マルチプロセッサシステムで共通に使用している装置については、「James Laudon, et. al. System Overview of the SGI Origin200/2000 Product Line. Proceeding of the 47th IEEE COMPUTER SOCIETY INTERNATIONAL CONFERENCE, pp. 150-156, Feb. 1997」に記載されている。Origin200/2000は、1枚以上のノードから構成され、各ノードにはCPU2個、主記憶、ディレクトリ、Hub chipから構成される。Hub chipは、CPUとの通信インタフェース制御部、主記憶及びディレクトリの通信インタフェース制御部、外部入出力インタフェース制御部、及びこれらのインタフェース制御部を結合するクロスバスイッチで構成される。小規模マルチプロセッサシステムに該当するOrigin200は、通常1枚のノードで構成されるが、2枚のノードをHub chipからの外部入出力インタフェースで直接接続した構成もある。大規模マルチプロセッサシステムに相当するOrigin2000では、2枚以上のノードボードをクロスバスイッチを搭載したルーターボードで接続している。以下の記述ではOrigin200とOrigin2000は区別せずOriginと呼ぶことにする。Originの例のように、システムの規模に関わらず同一のノードを使用して多様なシステム構成が可能なことは、開発コストの削減、開発期間の短縮に有効な手段である。

【0003】また、OriginはccNUMA (cache-coherent Non-Uniform Memory Access) 方式のマルチプロセッサで、ディレクトリ方式のキャッシュコヒーレンシ制御を行っている。この制御についての詳細は、「James Laudon, et. al. The SGI Origin: A ccNUMA Highly Scalable Server. Proceeding of 24th Annual Symposium on Computer Architecture, pp. 2

41-251, Jun. 1997」に記載されている。Originのメモリアクセスは通常以下に行われる。CPUで発行されたメモリアクセス要求は、要求するアドレスに対応する主記憶の存在するノードに転送され、該ノード内のディレクトリを検索する。ディレクトリは、要求したアドレスに対応するキャッシュラインごとに設けられ、どのノードのキャッシュメモリに転送され、どのような状態になっているか記録されている。ディレクトリを検索した結果、判明したノードのキャッシュメモリ、あるいは主記憶から読み出したデータは、要求発行元のCPUに転送される。

【0004】さらに、従来技術として、特開平9-138782に、キャッシュコヒーレンシ制御の為に、メモリアクセス要求の処理順序を決定する調停回路を有するクロスバスイッチが示されている。一般に、クロスバスイッチは、データの並列転送を持ち、スループット性能がバスより高いことは一般に知られている。しかし、メモリアクセスの順番が部分的に逆転し、キャッシュコヒーレンシを崩す可能性がある。Originではディレクトリ方式でキャッシュコヒーレンシを維持しているが、特開平9-138782で示されているマルチプロセッサシステムでは、CPUから発行されたメモリアクセス要求を、クロスバスイッチモジュール内の論理的に唯一の動作をする調停回路で一意の順番付けを行うことで、キャッシュコヒーレンシ制御を行っている。順番付けをされた要求は、クロスバスイッチモジュール内の選択回路を通して、CPUモジュールや主記憶モジュール、及び入出力モジュールに転送される。このように、クロスバスイッチにメモリアクセス要求の順番付け機能をもたせ、全てのCPUモジュールにメモリアクセス要求をブロードキャストすることでスヌープキャッシュを実現する方式を、以下においては擬似スヌープ方式と呼ぶことにする。また、擬似スヌープ方式を実現するためのメモリアクセス要求の順番付け機能を搭載したクロスバスイッチモジュールを擬似スヌープ機能付クロスバスイッチモジュールと呼ぶことにする。しかしながら、特開平9-138782のマルチプロセッサシステムでは、Originのように、大規模システムに用いられているノードだけを少数使用し、これらのノード間を直結して小規模システムを構成できるような装置にはなっていない。

【0005】

【発明が解決しようとする課題】上述のディレクトリ方式のキャッシュコヒーレンシ制御を行うマルチプロセッサシステムの場合、一般に、ディレクトリを参照を行う分だけLSI間転送の回数が多くなり、メモリレイテンシが増大するという課題がある。また、主記憶の量を増大する場合、ディレクトリの量も増大する。したがって、大容量主記憶を搭載する場合、大量のディレクトリ用のメモリが必要となるためコストが高くなる課題もある。

本発明の目的は、上述のOriginと同様に、ノード数の少ない小規模システムとノード数の多い大規模システムで共通のノードを使うことによる開発コストを削減できるマルチプロセッサシステムを提供することにある。本発明の他の目的は、小規模システムでは複数ノードを直結してシステムを構成し、ノード外部のクロスバスイッチを省略できるマルチプロセッサシステムを提供することにある。さらに、本発明の他の目的は、ディレクトリを用いないキャッシュコヒーレンシ制御方式を使うことで、ディレクトリ参照のオーバーヘッドによるメモリレイテンシの増分を削減し、かつ主記憶を増加させても、主記憶以外のコストが増加しないマルチプロセッサシステムを提供することにある。

【0006】

【課題を解決するための手段】上記目的を達成するため、本発明は、複数のノードを有する主記憶共有型マルチプロセッサシステムにおいて、該各ノードは、キャッシュメモリを有する1個以上のCPUモジュールと、1個以上の主記憶モジュールと、これらのモジュールと他ノードの間の通信制御を行うノード制御モジュールを備え、該ノード制御モジュールは、ノード間通信のインタフェースを制御する通信制御モジュールと、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンシ制御を行うクロスバスイッチを有するクロスバスイッチモジュールを有し、該クロスバスイッチモジュールは、モードレジスタを有し、該モードレジスタにセットされた値に応じて該クロスバスイッチを有効、あるいは無効にするようにしている。

【0007】また、前記クロスバスイッチモジュールは、前記モードレジスタに代えてモード信号ピンを有し、該モード信号ピンの信号値に応じて該クロスバスイッチを有効、あるいは無効にするようにしている。

【0008】また、前記クロスバスイッチモジュールは、全ノードから前記クロスバスイッチに転送されるメモリアクセス要求がどのノード内の主記憶モジュールに割り付けられているアドレス空間に対する要求であるかを判別する手段を有するようにしている。

【0009】また、前記ノードの外部に、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンシ制御を行うクロスバスイッチを有する外部クロスバスイッチモジュールを設け、前記各ノードを該外部クロスバスイッチモジュールに接続し、前記各ノード内のクロスバスイッチを無効にするようにしている。

【0010】また、前記ノード間を直接接続し、少なくとも1つのノード内のクロスバスイッチを有効にするようにしている。

【0011】

【発明の実施の形態】以下、本発明に係わるマルチプロセッサを図面に示したいくつかの実施例を参照してさら

に詳細に説明する。

《実施例1》

(装置構成の概略) 図1は、本発明に係わる主記憶共有型マルチプロセッサシステムの装置構成の概略を示す。図1において、 $5-i$ ($0 \leq i \leq N$, N は1以上の自然数)はノードであり、複数のノード間を信号線 $c-i$ 、 $d-i$ で結合する。次にノード $5-i$ の構成について説明する。各ノードは全て同一構造をしている。ノード $5-i$ は、CPUモジュール $4-i-j$ ($1 \leq j \leq k$, k は1以上の自然数)、キャッシュモジュール $6-i$ 、主記憶モジュール $7-i$ 、入出力モジュール $8-i$ 、ノード制御モジュール $2-i$ から構成される。ノード制御モジュール $2-i$ は、通信制御モジュール $3-i$ 、クロスバスイッチモジュール $1-i$ から構成されている。

【0012】各CPUモジュール $4-i-j$ は、ライトバック方式のプロセッサキャッシュ(図示せず)を有する。また、各CPUモジュールはプロセッサバス $h-i$ で結合されている。キャッシュモジュール $6-i$ は、ライトバック方式のキャッシュメモリ(図示せず)とキャッシュメモリ制御部(図示せず)で構成されている。キャッシュモジュール $6-i$ のキャッシュメモリは、ノード $5-i$ の全CPUモジュール $4-i-j$ で共有されている。本実施例のマルチプロセッサシステムのメモリ階層構造においては、プロセッサキャッシュと主記憶の中間に位置するキャッシュである。主記憶モジュール $7-i$ は全ノード $5-i$ で共有される主記憶空間の一部を構成している。入出力モジュール $8-i$ は複数の入出力装置、例えば、ディスク装置 $9-i$ に接続されている。この入出力モジュールには、他の入出力装置、例えば、回線接続装置(図示せず)等も接続されている。

【0013】通信制御モジュール $3-i$ は、CPUモジュール $4-i-j$ 、キャッシュモジュール $6-i$ 、主記憶モジュール $7-i$ 、及び入出力モジュール $8-i$ と、それぞれ、プロセッサバス $h-i$ 、信号線 $e-i$ 、 $f-i$ 、 $g-i$ で結合している。さらに、通信制御モジュール $3-i$ は、クロスバスイッチモジュール $1-i$ と、信号線 $a-i$ 、 $b-i$ で結合している。通信制御モジュール $3-i$ (内部の構造は図示せず)は、各モジュール間の通信制御を行っている。例えば、CPUモジュール $4-i-j$ からプロセッサバス $h-i$ に転送されたメモリアクセス要求をキャッシュモジュール $6-i$ に転送する処理や、他ノードに発行したメモリアクセス要求をクロスバスイッチモジュールを通して転送するための、通信プロトコル制御や、通信を行うデータのフォーマット変更等を行っている。クロスバスイッチモジュール $1-i$ は、自ノード内の通信制御モジュール $3-i$ と信号線 $a-i$ 、 $b-i$ で結合し、他ノードと信号線 $c-i$ 、 $d-i$ で結合している。

【0014】本発明では、同一構成のノードを、小規模構成のマルチプロセッサシステム、大規模構成のマルチ

プロセッサシステムで利用できることを特徴とする。ここで、マルチプロセッサシステムの規模とは、CPUモジュールの個数、つまり、ノードの枚数が少ないシステムを小規模、ノードの枚数が多いものを大規模と言っている。小規模、大規模の明確な数での切り分けはなく、実装上の都合でクロスバスイッチモジュール $1-i$ に設けられている他ノードとの結合信号線 $c-i$ 、 $d-i$ の組数によって決まる。図1は、小規模システムの構成を示し、図2は大規模システムの構成を示している。大規模構成システムの構成は、図1と同じ構成の複数ノード $5-i$ を、ノード $5-i$ の外に設けた外部クロスバスイッチモジュール 20 と、信号線 $c-0$ 、 \dots 、 $c-N$ 、 $d-0$ 、 \dots 、 $d-N$ で結合している。外部クロスバスイッチモジュール 20 は、前述した特開平9-138782号公報に記載されたものと同様のものである。

【0015】図3は、ノード制御モジュール $2-0$ の内部を示している。ノードは全て同一構造なのでここでは図1のノード $5-0$ を例にとり図3を用いて説明する。ノード制御モジュール $2-0$ は、既に説明したとおり通信制御モジュール $3-0$ と、クロスバスイッチモジュール $1-0$ を有している。クロスバスイッチモジュール $1-0$ は、擬似スヌープ機能付クロスバスイッチ 100 、このクロスバスイッチ 100 を使用する有効モードか使用しない無効モードかを設定するモードレジスタであるクロスババイパスフラグレジスタ 300 (説明及び図面においては、クロスババイパスフラグレジスタ 300 をクロスババイパスフラグ 300 とも記載する)、マルチプレクサ 303 、 304 、 $30-v$ ($2 \leq v \leq N$)から構成されている。ただし、ノード数が2枚の場合は、マルチプレクサ $30-2$ 、 \dots 、 $30-N$ 、信号線 $31-2$ 、 \dots 、 $31-N$ 、 $c-2$ 、 \dots 、 $c-N$ 、 $d-2$ 、 \dots 、 $d-N$ は使用されない。

【0016】クロスババイパスフラグレジスタ 300 は、1ビットのフラグレジスタで信号線 301 で通信制御モジュール $3-0$ と接続、信号線 302 でマルチプレクサ 303 、 304 、 $30-v$ ($2 \leq v \leq N$)と接続されている。マルチプレクサ 304 は、擬似クロスバからの信号 306 と別ノードからの信号線 $c-1$ を、クロスババイパスフラグからの信号線 302 を使ってどちらか一方を選択し、信号線 $b-0$ に出力している。マルチプレクサ 303 は、擬似クロスバからの信号 305 と通信制御モジュール $3-0$ からの信号線 $a-0$ を、クロスババイパスフラグからの信号線 302 を使ってどちらか一方を選択し、信号線 $d-1$ に出力している。

【0017】クロスババイパスフラグ 300 が1のとき、マルチプレクサ 303 、 304 共に、それぞれ信号線 $a-0$ 、 $c-1$ を選択し、擬似スヌープ機能付クロスバスイッチ 100 をバイパスする。クロスババイパスフラグ 300 が0のときは、マルチプレクサ 303 、 304 はそれぞれ、信号線 305 、 306 を選択する。

【0018】マルチプレクサ30-vは、クロスババイパスフラグが0のとき、他ノードからの信号線c-vを選択し、信号線31-vで擬似スヌープ機能付クロスバスイッチ100に接続する。クロスババイパスフラグ300が1のときは、信号線310を選択し、トランザクションが存在しないことを示すNOP信号311をマルチプレクサ30-vへ転送する。例えばオールゼロの信号がトランザクションが無い状態を示すならば、該NOP信号は配線時に固定値オールゼロに設定される。

【0019】このように、クロスババイパスフラグ300を0にした場合を「擬似スヌープ機能付クロスバスイッチ100が有効」、1にした場合を「擬似スヌープ機能付クロスバスイッチ100が無効」と呼ぶことにする。図3の擬似スヌープ機能付クロスバスイッチ100が無効の場合は、通信制御モジュール3-0の信号線a-0、b-0がノードの入出力として直接使われる。

本実施例1のマルチプロセッサシステムにおいて、擬似スヌープ機能付クロスバスイッチ100は、1枚のノードでのみ有効にし、他のノードは全て無効にする。図2の外部クロスバスイッチモジュール20を使った大規模構成の場合には、図3の擬似スヌープ機能付クロスバスイッチ100は、全ノードで無効にする。

【0020】図3のクロスババイパスフラグレジスタ300は、アドレス空間に配置したマップレジスタにしておき、システム起動時に一枚のノードのクロスババイパスフラグレジスタだけ0にしておき、残りのノードのクロスババイパスフラグレジスタは全て1になるように設定する。または、図4のように、モード信号ピン(クロスババイパスモード信号ピン)40を設置し、ボードの配線時にあらかじめ0か1に決めて配線してしまう方法や、ジャンプスイッチにしてあらかじめ手動で0か1に設定する方法(図示せず)を使用してもよい。

【0021】(擬似スヌープ機構付クロスバスイッチの機能説明) 擬似スヌープ機構付クロスバスイッチは、キャッシュコヒーレンシ制御のために、特開平9-138782に示されているメモリアクセス要求の順番付け機能を有している。クロスバスイッチは、データの並列転送ができることが特徴であることは一般に知られている。しかし、メモリアクセスの順番が部分的に逆転してしまい、キャッシュコヒーレンシを保証できない場合が起きうる。キャッシュコヒーレンシ制御方式としてバススヌープ方式がよく知られている。このバススヌープ方式では、バスに接続された複数のCPUがCPUバスにメモリアクセス要求を出すために、バス制御部にはバスの使用権を調停する機能がある。バスの調停で決まったメモリアクセス要求の順番が、システム全体で、つまりどのCPUでも一意であることでキャッシュの一貫性を保っている。このような、システムで一意のメモリアクセス要求の順番を決定する機能をクロスバスイッチに搭載することで擬似スヌープ機構付クロスバスイッチはキ

ャッシュコヒーレンシ制御を行っている。

【0022】(装置動作) ここでは、CPUモジュールで発行されたメモリアクセス要求が、全CPUモジュール、及び主記憶モジュールにブロードキャストされる動作を説明する。なお、以降では、メモリアクセスの要求や、それに対する応答等、プロセッサバスやクロスバスイッチなどで転送されるデータをトランザクションと呼ぶことにする。

【0023】図1のCPUモジュール4-0-1において、CPUが内蔵のキャッシュメモリに所望のデータが存在せず、キャッシュミスを起こしたとする。この場合、CPUモジュール4-0-1はプロセッサバスh-0にキャッシュミスしたアドレスのデータをアクセスするトランザクションを発行する。該トランザクションは、メモリの読出しである等のトランザクションの種類の情報、トランザクションを発行したCPUモジュール番号等の制御情報、要求するアドレスで構成されている(図示せず)。プロセッサバスh-0に転送されたトランザクションは、ノード制御モジュール2-0内の通信制御モジュール3-0を通して、キャッシュモジュール6-0に転送する。このキャッシュモジュール6-0で、該トランザクションの要求するアドレスのデータが保持されていない場合、所望のデータが他のノードでキャッシュされているかどうか調べるために全ノードにこのトランザクションを送ろうとする。この場合、通信制御モジュール3-0を通して、クロスバスイッチモジュール1-0に転送される。

【0024】本動作例では、ノード5-0のクロスババイパスフラグ300が0でクロスバスイッチが有効となっている。他のノード5-i ($1 \leq i \leq N$) ではクロスババイパスフラグ300が1となりクロスバスイッチが無効となっている。該クロスバスイッチモジュール1-0に入力されたトランザクションは信号線a-0を通して擬似スヌープ機能付クロスバスイッチ100に転送される。

【0025】図5は、図3に示されている擬似スヌープ機能付クロスバスイッチ100の構成を示している。入力信号線a-0、c-1、31-2、…、31-Nは、入力レジスタ500-x ($0 \leq x \leq N$)、520-x、521-xに接続している。ただし、図5では、全ての信号の図示はしていない、代表してa-0と31-Nのみを示している。TYPEレジスタ500-xには、クロスバスイッチに転送されるメモリアクセス要求、ならびにその応答データなど、トランザクションの種類を表す情報が格納される。MISCレジスタ520-xには、どこのノードで発行されたトランザクションであるか等、トランザクションの制御情報が格納される。ADDRESSレジスタ521-xには、要求しているアドレス情報が格納される。また、トランザクション全体は、信号線503-0を通し、トランザクションレジス

タ507-0に格納され、信号線t0を通して選択回路512-xに転送され、調停完了を待つ。

【0026】トランザクションの種類を表すTYPEレジスタ500-0のデータは信号線502-0を通してリクエスト制御504-0へ転送される。そして、レジスタ506-0の中の転送対象のノードに対応する要素に調停要求信号を書き込む。ここでは、キャッシュを持ったモジュールがあるノード、あるいは、主記憶が転送先となるので全ノードが転送対象となる。このレジスタ506-0から調停要求信号r00、…、r0Nが調停回路510-0、…、510-Nに転送される。

【0027】全ての調停回路が同じ調停動作を同時に行っている場合は、誤った順番付けを行うことはない。しかし、何らかの理由で出力レジスタ514-x、515-x、516-xにセットできず、各調停回路で調停を行う時間が異なる場合がありうるが、全調停回路は全て同じ優先順位で調停を行っているので、全て同じ順序でトランザクションが出力されていけばキャッシュコヒーレンシが崩れることはない。

【0028】調停回路510-x ($0 \leq x \leq N$) で調停に勝利した場合、調停完了信号g0xがリクエスト制御504-xに転送される。どのノードからのトランザクションが調停で勝ったかという情報は、信号線511-xを通し選択回路512-xに伝えられる。いま、調停要求r00が調停に勝利したとする。調停回路510-xは、リクエスト制御504-0に調停完了信号g0xを送り、同時に、選択回路512-0に信号線511-0を通して通知する。選択回路512-xは、調停に勝ったノードを選択し、トランザクションを信号線t0、及び信号線513-xを通し、出力レジスタ514-x、515-x、516-xに転送する。リクエスト制御504-xは、調停完了信号g0xが到着した後、調停要求の入ったレジスタ506-0をリセットし、同時に信号線505-0を通してトランザクションレジスタ505-0から削除を指示する。調停要求コマンドを格納するレジスタ506-0、及びトランザクションレジスタ505-0はリセットされ、次の入力待つ。出力レジスタのトランザクションは、図5の信号線306、…、d-N（正しくは、図3の306、305、d-2、…、d-N）を通して各ノードに転送される。

【0029】図1において、調停後のトランザクションは、ノード5-0ではクロスバスイッチモジュール1-0から信号線b-0を通し通信制御モジュールを通過し、キャッシュモジュール6-0、プロセッサバスh-0に転送され、各々のキャッシュタグを検索する。ノード5-1、…、5-Nでは、入力されたトランザクションは、クロスバスイッチをバイパスして通信制御モジュールへ入力され、同様にキャッシュモジュール6-1、…、6-N、プロセッサバスh-1、…、h-Nに転送され、各々のキャッシュタグを検索する。キャッシュの

状態を調査した後、各々のノードはこの調査結果をトランザクション発行元に報告し（この手段は図示していないが専用線を設けたり、メモリアクセス要求の転送で用いたクロスバスイッチを使ってもかまわない）、主記憶、あるいはキャッシュメモリからデータがトランザクション発行元に転送され、メモリアクセス要求は完了する。このデータ転送手段は、クロスバスイッチモジュールを併用する場合もあるし、別のデータ専用の通信路を用意してもよい。

【0030】《実施例2》

（装置構成の概略）図6は、本発明に係る主記憶共有型マルチプロセッサシステムの第2の実施例の装置構成の概略を示す。図6において、5-i ($0 \leq i \leq N$, Nは1以上の自然数)はノードであり、複数のノード間を信号線60-i-j、($0 \leq j \leq N$)で結合する。次にノード5-iの構成について説明する。各ノードは全て同一構造を有している。ノード5-iは、CPUモジュール4-i-j ($1 \leq j \leq k$, kは1以上の自然数)、キャッシュモジュール6-i、主記憶モジュール7-i、入出力モジュール8-i、ノード制御モジュール2-iから構成される。ノード制御モジュール2-iは、通信制御モジュール3-i、クロスバスイッチモジュール1-iから構成されている。

【0031】各CPUモジュール4-i-jは、ライトバック方式のプロセッサキャッシュ（図示せず）を有する。また、各CPUモジュールはプロセッサバスh-iで結合されている。キャッシュモジュール6-iは、ライトバック方式のキャッシュメモリ（図示せず）とキャッシュメモリ制御部（図示せず）で構成されている。キャッシュモジュール6-iのキャッシュメモリは、ノード5-iの全CPUモジュール4-i-jで共有されている。本実施例のマルチプロセッサシステムのメモリ階層構造においては、プロセッサキャッシュと主記憶の中間に位置するキャッシュである。主記憶モジュール7-iは全ノード5-iで共有される主記憶空間の一部を構成している。入出力モジュール8-iは複数の入出力装置、例えば、ディスク装置9-iに接続されている。この入出力モジュールには、他の入出力装置、例えば、回線接続装置（図示せず）等も接続されている。

【0032】通信制御モジュール3-iは、CPUモジュール4-i-j、キャッシュモジュール6-i、主記憶モジュール7-i、入出力モジュール8-iを、それぞれ、プロセッサバスh-i、信号線e-i、f-i、g-iと結合し、クロスバスイッチモジュール1-iとは、信号線a-i、b-iで結合している。通信制御モジュール3-i（内部の構造は図示せず）は、各モジュール間の通信制御を行っている。通信制御の内容は、CPUモジュール4-i-jからプロセッサバスh-iに転送されたメモリアクセス要求をキャッシュモジュール6-iに転送する処理や、他ノードに発行したメモリ

アクセス要求をクロスバスイッチモジュールを通して転送するための通信プロトコル制御や、通信を行うデータのフォーマット変更等である。

【0033】クロスバスイッチモジュール1-iは、自ノード内の通信制御モジュール3-iと信号線a-i、b-iで結合し、他ノードと信号線60-i-jで結合している。図7に示すアドレスマップ700は、主記憶空間のアドレスと、どのノードの主記憶に割り付けられているかアドレスとノードの対応表である。アドレスマップ700はメモリマップドレジスタとして主記憶空間に配置され、システムが立ち上がる時、ノードに割り振られているメモリ空間のアドレスをセットする。また、ノード上のジャンパスイッチなどを使用してアドレスマップに格納するアドレス情報を定義する方法もある。

【0034】図7は、ノード制御モジュール2-0の内部を示している。ノードは全て同一構造で、同一動作を行うので、本実施例では図6のノード5-0の動作のみを図7を使って説明する。ノード制御モジュールは、既に説明したとおり通信制御モジュール3-0と、クロスバスイッチモジュール1-0を有している。クロスバスイッチモジュール1-0は、擬似スヌープ機能付クロスバスイッチ100、トランザクション判別部70-0、アドレスマップ700、マルチプレクサ704、705、クロスババイパスフラグレジスタ710（説明及び図面においては、クロスババイパスフラグレジスタ710をクロスババイパスフラグ710とも記載する）から構成されている。クロスババイパスフラグは、実施例1の図3に示されているものと同様の働きをし、図7のクロスババイパスフラグ710を1に設定すると、通信制御モジュール3-0と擬似スヌープ機能付クロスバスイッチ間の信号線a-0、b-0がノードの入出力信号60-01、60-10と直結する。また、クロスババイパスフラグが0の時は、通信制御モジュールの信号線a-0はトランザクション判別部70-0に、信号線b-0は擬似スヌープ機能付クロスバスイッチ100からの出力信号となる。

【0035】クロスババイパスフラグは、ノード間を直結する小規模システム時には、全ノードで0に設定し、外部クロスバスイッチモジュールでノードを結合する大規模構成時には1に設定する。なお、クロスババイパスフラグ710は信号線711で通信制御モジュール3-0と接続し、実施例1と同様にメモリマップドレジスタにして、システム起動時に設定するようにする。あるいは、実施例1と同様にクロスババイパスフラグレジスタの代わりにモード信号ピン（図示せず）を設ける方法でもよい。

【0036】トランザクション判別部70-0は、通信制御モジュールからの出力信号線a-0とアドレスマップ700からの信号線702を入力とし、擬似スヌープ機能付クロスバスイッチ100と接続する信号線71-

0を出力する。擬似スヌープ機能付クロスバスイッチ100の擬似スヌープ機能については実施例1で説明した通りである。図7のクロスバスイッチモジュール1-0内の擬似スヌープ機能付クロスバスイッチ100は、図12に示す表に基づいた動作を行う。図12に基づく動作をフローチャートにしたものが図9である。

【0037】図8において、トランザクション判別部70-iは、入力レジスタ801、802、803、804、805、及び806、要求ノード検索回路800から構成されている。入力レジスタは801、802、803、804、805、及び806は信号線a-iと接続している。要求ノード検索回路800は入力として、入力レジスタ804と信号線807で、及びアドレスマップと信号線702で接続されている。要求ノード検索回路800の出力信号で、AMレジスタ809の内容を変更する。該AMレジスタには、ノード番号が格納されている。このノード番号は、トランザクションが要求しているアドレスがどのノードの主記憶に割り付けられているか示している。

【0038】（装置動作）ここでは、実施例1と同様にCPUモジュールで発行されたメモリアクセス要求を行うトランザクションが、全CPUモジュール、及び主記憶モジュールにブロードキャストされる動作を説明する。本実施例では、全ノードの主記憶上のデータ、あるいは他ノードのキャッシュにアクセスを行う場合を説明する。本実施例では、メモリアクセス要求を行うトランザクション発行元が自ノードの主記憶に割り付けられたアドレスをアクセスする場合と他ノードの主記憶に割り付けられたアドレスをアクセスする場合を示す。

【0039】まず、動作例（1）として自ノードの主記憶に割り付けられたアドレスをアクセスする場合について説明する。図6のCPUモジュール4-0-1において、CPUが内蔵のキャッシュメモリに所望のデータが存在せず、キャッシュミスを起こしたとする。この場合、CPUモジュール4-0-1はプロセッサバスh-0にキャッシュミスしたアドレスのデータをアクセスするトランザクションを発行する。該トランザクションは、メモリの読出しである等のトランザクションの種類の情報、トランザクションを発行したCPUモジュール番号等の制御情報、要求するアドレスで構成されている（図示せず）。プロセッサバスh-0に転送されたトランザクションは、ノード制御モジュール2-0内の通信制御モジュール3-0を通して、キャッシュモジュール6-0に転送する。該キャッシュモジュールで、当該トランザクションの要求するアドレスのデータが保持されていない場合、所望のデータが他のノードでキャッシュされているかどうか調べるために全ノードにこのトランザクションを送ろうとする。この場合、通信制御モジュール3-0から信号線a-0を通して、クロスバスイッチモジュール1-0に転送される。

【0040】クロスバススイッチモジュール1-0に転送されたトランザクションは、図8のトランザクション判別部70-iに転送される。まず、入力レジスタ801、802、803、804、805、806にトランザクションを格納する。TYPEレジスタ802はトランザクションの種類を示す情報が格納される。MISCレジスタ803には、トランザクションのID番号等の制御情報が格納されている。ADDRESSレジスタ804には、トランザクションのアクセスするアドレスが、NIDレジスタ813には、トランザクションを発行したノードの番号を示す情報が格納されている。また、AMレジスタ806には、トランザクションが要求するアドレスが、どのノードの主記憶に割り付けられているか、そのノード番号が格納されている。ARBレジスタ801は、1ビットのレジスタで当該トランザクションがクロスバススイッチモジュールで既に調停を終了している場合は1、調停未完了のときは0が格納されている。本ケースでは、まだクロスバススイッチで調停が行われていないのでARB=0となっている。要求ノード検索回路800は、アクセス対象が自ノードに設定されたメモリ空間なのでAMレジスタに自ノード番号である'0'を書き込む。ただし、本実施例では、図6のノード5-i ($0 \leq i \leq N$)のNIDをiとしている。

【0041】図10の疑似スヌープ機能付クロスバススイッチ100に入力されたトランザクションは、入力レジスタであるTYPEレジスタ500-0、MISCレジスタ520-0、ADDRESSレジスタ521-0、ARBレジスタ1100-0、NIDレジスタ1102-0、AMレジスタ1103-0に格納される。レジスタ1100-0、500-0、520-0、1102-0、1103-0はコマンド信号としてリクエスト制御504-0に送られ、送付ノードの番号のついたレジスタ506-0に調停要求信号をセットする。また、トランザクション全体は信号線503-0を通してトランザクションレジスタ507-0に格納される。

【0042】図9のフローチャートに示す判定をリクエスト制御504-i ($0 \leq i \leq N$)で行う。本ケースでは、ARB=0、AM=0なので図9の条件1001は満たさず、条件1003を満たすので、図9の処理1004を実行する。処理1004により、信号線505-0でARB=1にセットし、レジスタ506-0は全ノードを対象に調停要求をセットする。

【0043】リクエスト制御504-0から調停回路510-iに転送された調停要求r00、…、r0Nが調停回路で勝利すると選択回路512-0に勝利したノードの番号、ここでは0を伝え、信号線t0を通してトランザクションを出力レジスタ1101i、514-i、515-i、516-i、1004-i、1105-iにセットし、各ノードに転送する。他ノードには、図7の信号線60-0i ($1 \leq i \leq N$)を通して伝えられる。

自ノードの通信制御モジュール3-0にはb-0を通してトランザクションが伝えられる。

【0044】トランザクションが他ノードに到着した場合を説明する、ここでは簡単の為、図6のノード5-1での場合を説明する。図6のノード5-0の信号線60-01を通してトランザクションがノード5-1に伝えられる。当該トランザクションは信号線60-01を通して図11の疑似スヌープ機能付クロスバススイッチ100に転送される。疑似スヌープ機能付クロスバススイッチ100では、図9のフローチャートに従い、ARB=1なので条件1001を満足するので1002の処理を行う。疑似スヌープ機能付クロスバススイッチ100の動作は先の説明とほとんど同様だが、図10のリクエスト制御504-1のレジスタ506-1で自ノードに対応するレジスタのみ調停要求をセットする。後の処理は、先ほどと同様の調停処理を行い、図11の信号線b-1を通して通信制御モジュール3-1にトランザクションを転送する。このようにして、全ノードの通信制御モジュールにトランザクションがブロードキャストされる。ブロードキャスト後の処理は実施例1の動作例と同様である。

【0045】次に、動作例(2)として他ノードの主記憶に割り付けられたアドレスをアクセスする場合について説明する。ここでは、図6のノード5-1に割り当てられた主記憶にアクセスすると仮定する。図6のCPUモジュール4-0-1で発行されたトランザクションがキャッシュミスを起こし、クロスバススイッチモジュール1-0に転送されるまでの動作は動作例(1)の場合と同様である。

【0046】図6のクロスバススイッチモジュールにトランザクションが転送された後、図10の疑似スヌープ機能付クロスバススイッチのリクエスト制御504-0では、図9の処理が行われる。本動作例では、ARB=0、AM=1なので、条件1001、1003は満たされず、処理1005が行われる。処理1005では、AMレジスタ1103-0で示されたノードにのみトランザクションを転送するので、図10のレジスタ506-0において、該AMレジスタで示されているノードのみに調停要求をセットする。調停後、該トランザクションは、図6のノード5-1に転送される。当該ノードの内蔵する疑似スヌープ機能付クロスバススイッチでは、図9の1001の条件は満たさず、条件1003を満たすので処理1004が行われる。この後の動作は動作例(1)と同様である。

【0047】

【発明の効果】本発明によれば、各ノードにクロスバススイッチを搭載し、直接ノード間結合することで小規模マルチプロセッサシステムを構成できる。しかも、このノードは大規模システムにも利用できるため開発コストを削減できる。さらに、本発明によれば、ディレクトリを

用いずクロスバスイッチでキャッシュコヒーレンシ制御をするため、ディレクトリを検索するための通信コストが削減されメモリレイテンシの削減が可能となる。また、主記憶を増加する場合、ディレクトリを使用していないため主記憶以外の資源を追加する必要がなく、低コストでシステムのスケールアップが可能となる。

【図面の簡単な説明】

【図1】第1の実施例に係わる主記憶共有型マルチプロセッサシステムの装置構成の概略を示す図である。

【図2】第1の実施例におけるノードを外部クロスバスイッチモジュールに接続した大規模マルチプロセッサシステムの全体構成を示す図である。

【図3】第1の実施例におけるノード制御モジュールの構成を示す図である。

【図4】図3に示すノード制御モジュールの変形例を示す図である。

【図5】第1の実施例における疑似スヌープ機能付クロスバスイッチの構成を示す図である。

【図6】第2の実施例に係わる主記憶共有型マルチプロセッサシステムの装置構成の概略を示す図である。

【図7】第2の実施例におけるノード制御モジュールの構成を示す図である。

【図8】第2の実施例におけるトランザクション判別部の構成を示す図である。

【図9】第2の実施例における疑似スヌープ機能付クロスバスイッチを構成するリクエスト制御の動作の一部のフローチャートを示す図である。

【図10】第2の実施例における疑似スヌープ機能付クロスバスイッチの構成を示す図である。

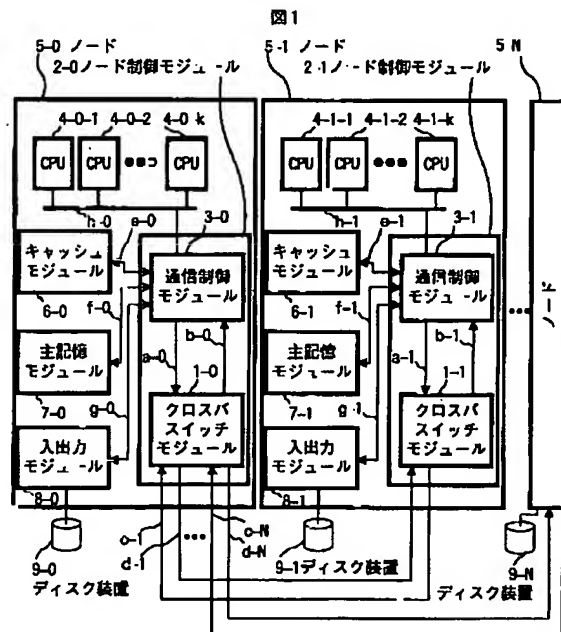
【図11】第2の実施例における他のノードのノード制御モジュールの構成を示す図である。

【図12】第2の実施例における疑似スヌープ機能付クロスバスイッチを構成するリクエスト制御の動作の一部を記述した表を示す図である。

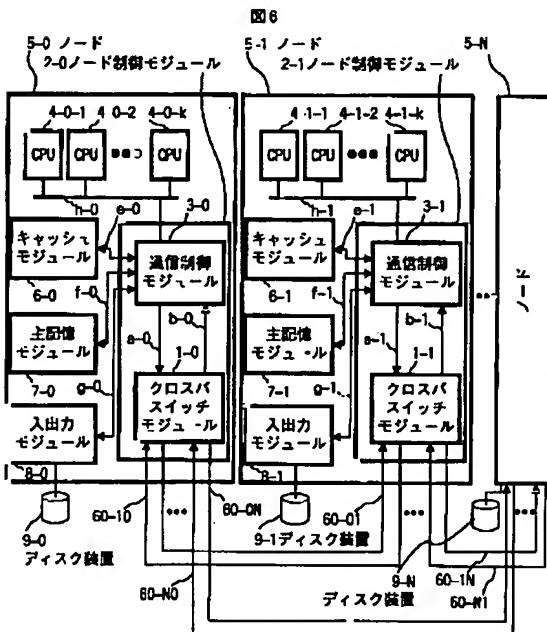
【符号の説明】

- 1-0, 1-1 クロスバスイッチモジュール
- 2-0, 2-1 ノード制御モジュール
- 3-1, 3-2 通信制御モジュール
- 4-0-1~4-1-k CPU
- 5-0, 5-1, 5-N ノード
- 6-0, 6-1 キャッシュモジュール
- 7-0, 7-1 主記憶モジュール
- 8-0, 8-1 入出力モジュール
- 9-0, 9-1, 9-N ディスク装置
- 20 外部クロスバスイッチモジュール
- 40 クロスババイパスモード信号ピン
- 70-0, 70-1 トランザクション判別部
- 100 疑似スヌープ機能付クロスバスイッチ
- 300, 710 クロスババイパスフラグレジスタ
- 303, 304, 30-2, 30-N, 704, 705 マルチプレクサ
- 700 アドレスマップ
- 800 要求ノード検索回路

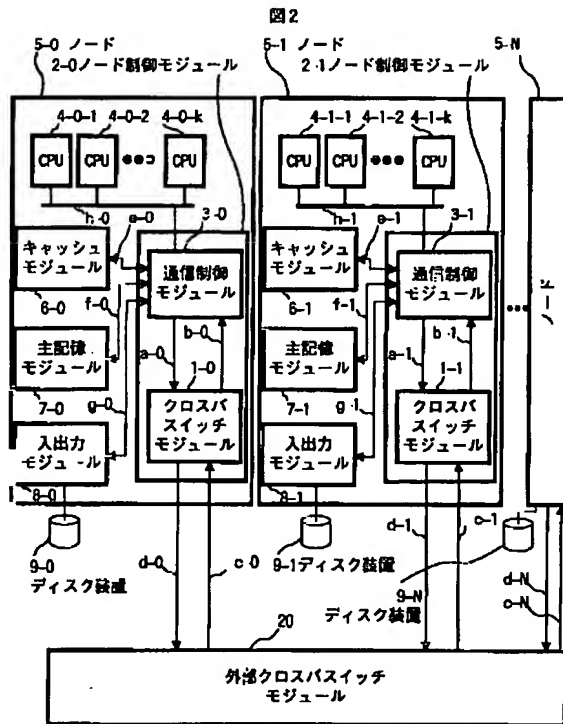
【図1】



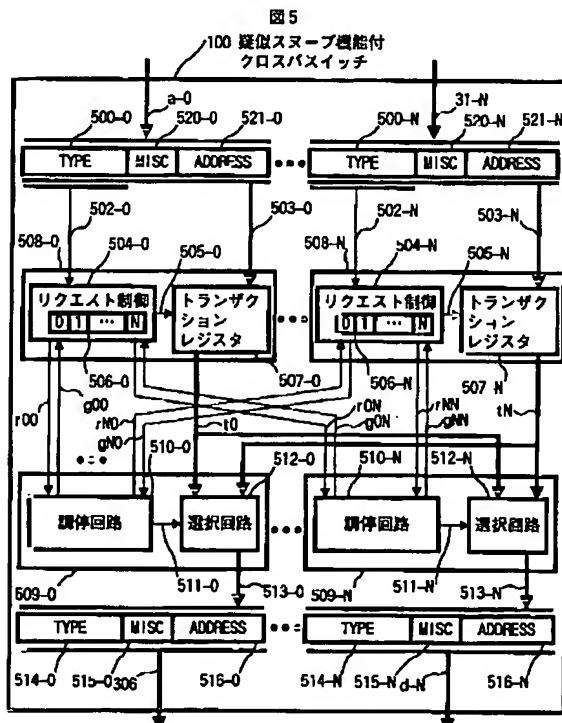
【図6】



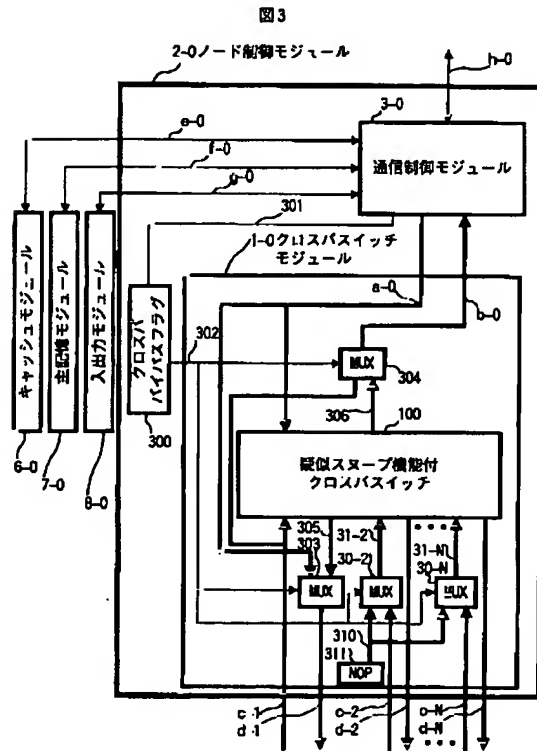
【図2】



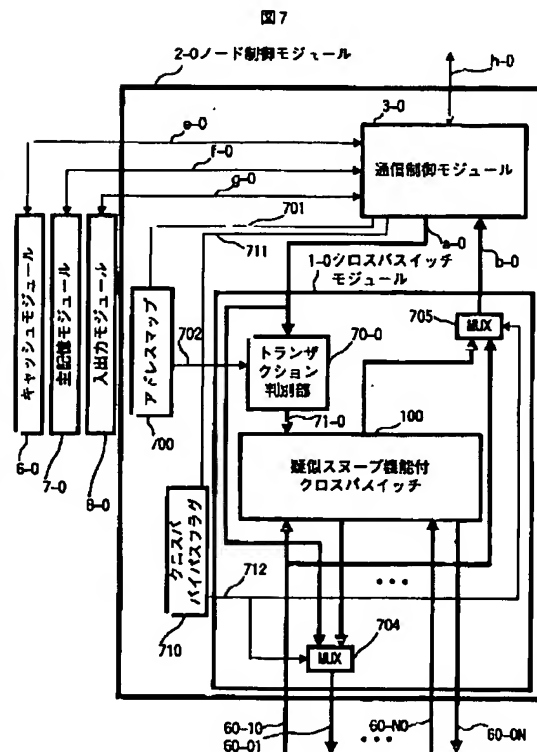
【図5】



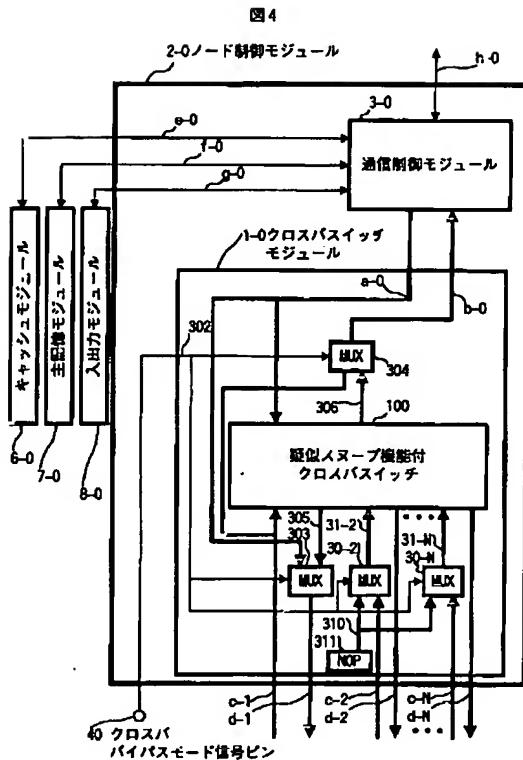
【図3】



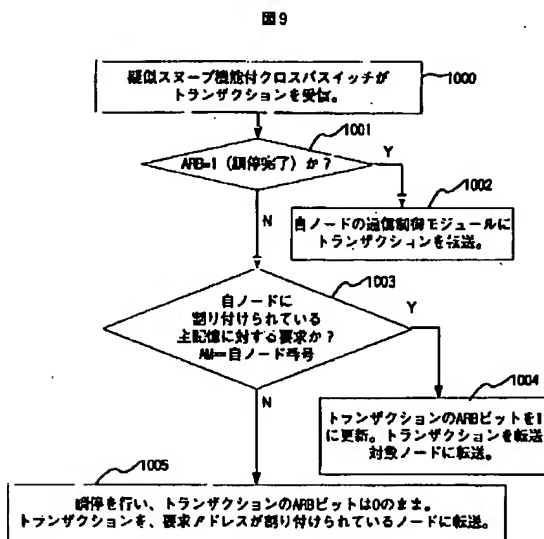
【図7】



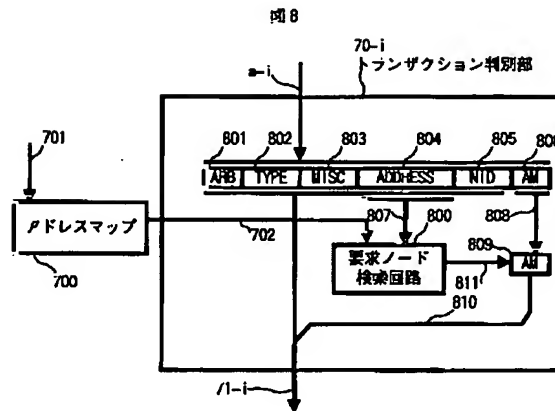
【図4】



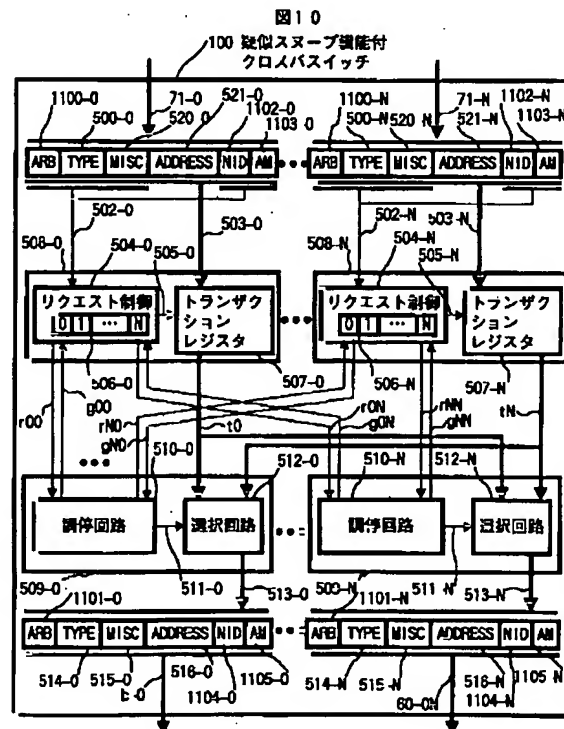
【図9】



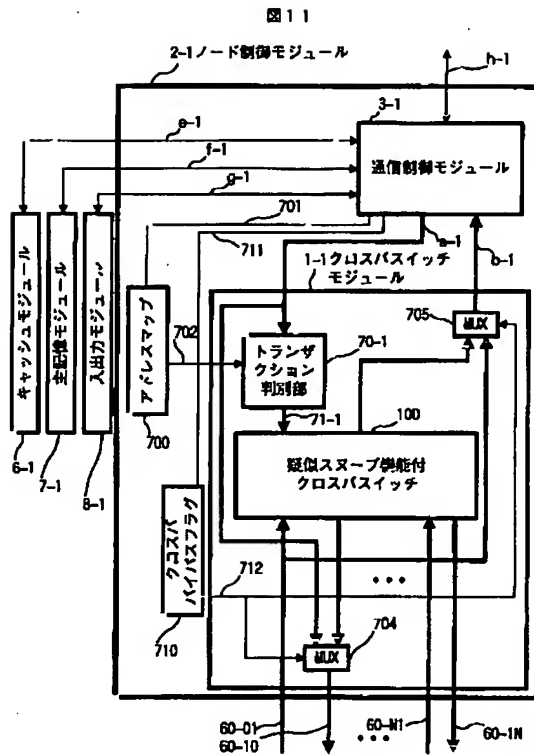
【図8】



【図10】



【図11】



【図12】

図12

		トランザクションが要求するアドレス空間	
		自ノードの主記憶	他ノードの主記憶
調停前 ARB=0	トランザクション発行元 自ノード	転送対象ノード に転送 ARB=1に更新	要求するアドレスが 割り当てられている ノードに転送、 ARBの更新はしない
	トランザクション発行元 他ノード		
調停後 ARB=1	トランザクション発行 任意のノード	自ノードの通信制御モジュール (最終的には、CPUモジュール、 キャッシュモジュール)に転送	

フロントページの続き

(72)発明者 對馬 雄次
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者 上原 敬太郎
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者 濱中 直樹
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者 庄内 亨
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内

Fターム(参考) 5B045 BB16 BB44 DD12